METHOD AND SYSTEM FOR UTTERANCE VERIFICATION

BACKGROUND OF THE INVENTION

1. Field of the Invention

The present invention relates to the technical field of utterance verification and, more particularly, to a method and a system for utterance verification which are adapted to a noisy environment.

2. Description of Related Art

Conventionally, in the field of data processing, an utterance verification technique is employed to verify the correctness of a candidate string which is obtained through conducting speech recognition on a speech segment. Therefore, correct action can be taken on according to the candidate string which is verified and regarded as the correct answer. For example, in voice dialing system, a voice input for a set of telephone number is requested, a digit-string will be recognized and verified for the input speech, and the recognized digit-string will be dialed if it is verified and regarded as the correct answer.

In many well-known utterance verification techniques, the most widely employed techniques are the decoding based approach and a hypothesis testing based approach.

FIG. 1 is a block diagram illustrating a decoding based utterance verification technique. As shown, a word "Hi", denoted by two phonetic symbols "h" and "ai", represented by an input speech 51 has been recognized. In most decoding based systems, "Hi" is taken as a unit for decoding and parameters used for decoding are calculated for the word "Hi".

Besides, more than one parameters 52are usually used in this approach, such as acoustic score 521, language model score 522, N-best information 523, etc included in the parameter set 52. A decoder 53 is then activated to combine these scores for obtaining a verification score 54. Finally, the verification score 54 is compared with a predetermined threshold to decide recognized word "Hi" should be accepted or not. This approach can be implemented by LDA (Linear Discriminative Analysis), decision tree analysis, or neural network. However, various and complicated parameters are required for this approach. It is an extra and time-consuming effort for many speech applications.

FIG. 2 is a block diagram illustrating a hypothesis testing based utterance verification technique. As shown, a word "Hi" represented by an input speech 61 has been recognized. And, the input speech is segmented into sub-word segments of "h" and "ai". Verification model 621 (623) and anti-model 622 (624) of sub-word "h" ("ai") are used to test the sub-word segment of "h" ("ai"). And the resulted log likelihood ratio is regarded as the test score 631 (632) of sub-word "h" ("ai"). The verification score 64 of "Hi" is obtained through combining test scores 631 and 632. Finally, the verification score 64 is compared with a predetermined threshold to decide whether the recognized word "Hi" should be accepted or not. However, the hypothesis testing based utterance verification technique requires verification model and anti-model for each sub-word. And, two tests are required for each sub-word segment. System load will be significant increased by using this approach.

Moreover, both the decoding and hypothesis testing approaches are only applicable to noiseless environment. Hence, verification performance will be degraded greatly in noisy environment. As a result, reliability of recognition system will be poor.

Therefore, it is desirable to provide novel method and system for utterance verification in order to mitigate and/or obviate the aforementioned problems.

SUMMARY OF THE INVENTION

An object of the present invention is to provide a method and a system for utterance verification both being adapted to a noisy environment so as to increase the reliability of speech recognition system.

Another object of the present invention is to provide a method and a system for utterance verification wherein a classifier is provided for each verification unit to decrease system load and increase system performance.

A further object of the present invention is to provide a method and a system for utterance verification both being easily integrated to various speech recognition systems.

In one aspect of the present invention there is to provide a method for utterance verification comprising the steps of extracting a sequence of feature vectors from an input speech; inputting the sequence of feature vectors to a speech recognizer for obtaining at least one candidate string; segmenting the input speech into at least one speech segment according to the content of candidate string, which comprises individual recognition units, wherein each speech segment corresponds to a recognition unit and

each recognition unit corresponds to a verification unit; generating a sequence of verification feature vectors according to the sequence of feature vectors of the speech segments normalized by the normalization parameters of the verification unit corresponding to the speech segments; utilizing a classifier for each speech segment to calculate verification scores corresponding to the speech segments according to the sequence of verification feature vectors corresponding to each speech segment; combining the verification scores of all speech segments for obtaining an utterance verification score of candidate string; and comparing the utterance verification score of candidate string with a predetermined threshold so as to accept the candidate string if the utterance verification score is larger than the predetermined threshold.

In another aspect of the present invention there is to provide a system for utterance verification comprising a feature vector extraction module for extracting a sequence of feature vectors from an input speech; a speech recognition module for obtaining at least one candidate string from the sequence of feature vectors; a speech segmentation module for segmenting the input speech into at least one speech segment according to the content of candidate string, which comprises individual recognition units, wherein each speech segment corresponds to a recognition unit and each recognition unit corresponds to a verification unit; a verification feature vector generation module for generating a sequence of verification feature vectors according to the sequence of feature vectors of the speech segments normalized by the normalization parameters of the verification unit

corresponding to the speech segments; a verification score calculation module for utilizing a classifier for each speech segment to calculate verification scores corresponding to the speech segments according to the sequence of verification feature vectors corresponding to each speech segment; a verification score combination module for combining the verification scores of all speech segments for obtaining an utterance verification score of candidate string; and a decision module for comparing the utterance verification score of candidate string with a predetermined threshold so as to accept the candidate string if the utterance verification score is larger than the predetermined threshold.

Other objects, advantages, and novel features of the invention will become more apparent from the detailed description when taken in conjunction with the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a block diagram illustrating a conventional decoding based utterance verification technique;

FIG. 2 is a block diagram illustrating a conventional hypothesis testing based utterance verification technique;

FIG. 3 is a block diagram illustrating a system for utterance verification according to the present invention;

FIG. 4 is a flow chart illustrating a process of utterance verification according to the present invention; and

FIG. 5 presents schematically a neural network according to the present invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

With reference to FIG. 3, there is shown a computer system 20 for implementing utterance verification in accordance with the invention. The computer system 20 is enabled to perform an utterance verification process on an input speech 10. The system 20 comprises a feature vector extraction module 21, a speech recognition module 22, an acoustic models 28, a vocabulary 29 having a plurality of words, a speech segmentation module 23, a verification feature vector generation module 24, a normalization database 11 having a plurality of normalization parameters, a verification score combination module 26, a decision module 27, and a verification score calculation module 25. The system 20 further comprises a plurality of neural networks 121, 122, and 123 ... etc. to be used for verification score calculation. In this embodiment, the neural networks are carried out by an MLP (multi-layer perceptron).

With reference to FIG. 4, there is shown a flow chart illustrating a process of utterance verification according to the invention. First, the input speech 10 is windowed into a plurality of frames in feature vector extraction module 21 for extracting a sequence of feature vectors 211 (step S401). Each feature vector is comprised of a plurality of dimensional values in the sequence of feature vectors 211. In this embodiment, the frame length is 160 points (i.e., 20ms). The overlapped length of frame is 80 points (i.e., 10ms). Also, the Hamming window is applied to smooth each data frame. Furthermore, each feature vector comprises 26 dimensional values including 12 Mel-ceptral coefficients, 12 delta-ceptral coefficients, a

delta-log-energy, and a delta-delta-log-energy. In this embodiment, a ceptral mean normalization is used for channel compensation.

The extracted sequence of feature vectors 211 is then inputted to the speech recognition module 22 for obtaining at lease one candidate string 221 (step S402). In the speech recognition module 22, the HMM (Hidden Markov model) based acoustic models 28 and a vocabulary 29 having a plurality of words are used in recognition for generating at least one candidate string 221. Next, the speech segmentation module 23 segments the input speech 10 into a plurality of speech segments 231, 232, and 233 according to the content of the candidate string 221 (step S403), where the content of candidate string 221 is represented by individual recognition units, and, each recognition unit corresponds to a verification unit. As a result, each of the speech segments 231, 232, and 233 corresponds to a verification unit. In a case that the candidate string 221 is an English word, each speech segment, 231, 232, and 233, can be corresponded to a verification unit. For example, "sky" will be segmented into three speech segments denoted by the verification unit of "s", "k", and "ai".

Then, the sequence of feature vectors 211 associated with the speech segments 231, 232, and 233 is sent to the verification feature vector generation module 24 for generating a sequence of verification feature vectors 241, 242, and 243. Note that the generation of the sequence of verification feature vectors 241, 242, and 243 is done by normalizing the sequence of feature vectors 211 according to the normalization parameters of verification unit corresponding to speech segment 231, 232, and 233.

The normalization parameters of the verification unit are the means and standard deviations of feature vectors through calculating feature vectors corresponding to the verification unit from training data in advance. The normalization parameters are then stored in the normalization database 11. Therefore, as to the speech segment 231, the sequence of verification feature vectors 241 is obtained by normalizing the sequence of feature vectors 211 corresponding to the speech segment 231 by means of normalization parameters of the verification unit corresponding to the speech segment 231 (step S404). Also, the sequence of verification feature vectors 241 comprises a plurality of dimensional values. Likewise, the sequence of verification feature vectors 242 and 243 corresponding to the speech segments 232, 233 can also be obtained respectively.

Next, the verification feature vectors 241, 242, and 243 are sequentially inputted to the verification score calculation module 25 for calculating verification scores 251, 252, and 253. In this process, the neural networks 121, 122, and 123 corresponding to the verification units are applied, where these verification units are associated with the speech segments 231, 232, and 233. Each of the neural networks 121, 122, and 123 is implemented by an MLP. Note that the calculation of the verification scores 251, 252, and 253 is done by performing a feed-forward processing on the neural networks 121, 122, and 123 using the verification feature vectors 241, 242, and 243, respectively (step S405). A description about calculating the verification score 251 using the neural network 121 is given hereinafter.

With reference to FIG. 5, it shows schematically the neural network

121. The neural network 121 is an MLP and comprises an input layer 31, a hidden layer 32, and an output layer 33. Also, each verification unit corresponds to a neural network. Input neurons 311 of the input layer 31 sequentially receive feature values D_i ($1 \leqq i \leqq N$) of each verification feature vector of the sequence of the verification feature vectors 241 and then output the feature values D_i to the hidden layer 32. Hidden neurons 321 of the hidden layer 32 sequentially receive all of the output values outputted from the input layer 31, calculate the corresponding result for each hidden neuron, and output these results to the output layer 33. The output layer 33 uses only one output neuron 331 to receive all of the output values outputted from the hidden layer 32, calculates a verification result, and outputs this verification result. Therefore, the verification results of the sequence of verification feature vectors 241 can be obtained according to the aforesaid feed-forward processing, and the verification score 251 is the mean of the verification results.

As to each output neuron of the neural network 121, the following equation is used to calculate out_j except for those with the input neuron directly used to output the feature value of the verification feature vector:

$$out_j = \frac{1}{1+\exp(-\sum_i w_{ji} out_i + b_j)},$$

wherein out_j is the output value of the j-th neuron of the present layer (hidden layer 32 or output layer 33), out_i is the output value of the i-th neuron of the former layer (input layer 31 or hidden layer 32), w_{ji} is a weight from the i-th neuron of the former layer to the j-th neuron of the

present layer, and b_j is a bias of the j-th neuron of the present layer.

After calculating all verification scores 251, 252, and 253, the verification score combination module 26 then calculates a mean of the verification scores 251, 252, and 253 to obtain an utterance verification score 261 of the candidate string 221 (step S406). Finally, the decision module 27 compares the utterance verification score 261 with a predetermined threshold (step S407). The system 20 will accept the candidate string 221 if the utterance verification score 261 is larger than the threshold (step S408); otherwise the candidate string 221 is rejected (step S409).

For enabling the MLP to verify whether the speech segment is the corresponding verification unit or not according to the aforesaid sequence of verification feature vectors in this embodiment, the training data of the MLP used for training a specific verification unit comprises the sequence of verification feature vectors of the speech segments corresponding to the verification unit and those not corresponding to the verification unit for enabling the MLP to learn the difference between each other. In the training process, a corresponding target value is firstly set for the sequence of the verification feature vectors of speech segments. Then, an error back-propagation algorithm is used to train the MLP. The weights and biases of the MLP can be adjusted by reducing the mean square error between the real verification score output and the desired verification score. For example, for training an MLP corresponding to a verification unit of "u", the training speech segments include the speech segments corresponding to

the verification unit of "u" and the speech segments that are not corresponding to the verification unit of "u". The desired verification score is 1 if the speech segment corresponds to the verification unit of "u". Otherwise, the desired verification score is 0. In such a manner, a real verification score output of the MLP will approximate to the target value by iteratively learning. Therefore, the MLP can verify whether the speech segment is the corresponding verification unit or not by using the sequence of verification feature vectors.

For enabling both the method and system of the invention to be adapted to a noisy environment, the training data used for training the MLPs are pre-corrupted by noise with different power levels of SNR (Signal to Noise Ratio). For example, the speech segments corrupted by in-car noise with SNRs of 9dB, 3dB, 0dB, -3dB, and -9dB are used to train the MLPs. As such, the MLP can learn voice characteristics of the speech segments under different levels of in-car noises. By utilizing this, both the method and system of the invention can determine whether a speech segment corrupted by noise is a corresponding verification unit or not.

Although the present invention has been explained in relation to its preferred embodiment, it is to be understood that many other possible modifications and variations can be made without departing from the spirit and scope of the invention as hereinafter claimed.